Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 777–787
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Term norm distribution and its effects on Latent Semantic Indexing

Parry Husbands *, Horst Simon, Chris Ding

Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Received 15 June 2003; accepted 24 March 2004
Available online 18 May 2004

Abstract

Latent Semantic Indexing (LSI) uses the singular value decomposition to reduce noisy dimensions and improve the performance of text retrieval systems. Preliminary results have shown modest improvements in retrieval accuracy and recall, but these have mainly explored *small* collections. In this paper we investigate text retrieval on a *larger* document collection (TREC) and focus on distribution of word *norm* (magnitude). Our results indicate the inadequacy of word representations in LSI space on large collections. We emphasize the query expansion interpretation of LSI and propose an LSI term normalization that achieves better performance on larger collections.

Keywords: Information retrieval; LSI; TREC

1. Introduction

The use of Latent Semantic Indexing (LSI) has been proposed for text retrieval in several recent works (Berry, Dumais, & O'Brien, 1995; Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Dumais, 1991; Hull, 1994). This technique uses the singular value decomposition (SVD) (Golub & Van Loan, 1996) to project very high dimensional document and query vectors into a low dimensional space. In this new space it is reasoned that the underlying structure of the collection is revealed thus enhancing retrieval performance. Furthermore, LSI can be alternatively reviewed as a query expansion method (see Sections 2.2 and 5), so that *recall* is generally improved. Experiments indicate both improved retrieval precision and recall when LSI is adopted (Ando & Lee, 2001; Bartell, Cottrell, & Belew, 1995; Berry et al., 1995; Deerwester et al., 1990; Dumais, 1991; Hull, 1994; Zha, Marques, & Simon, 1998). LSI also improves text categorization (Baker & McCallum, 1998; Dumais, 1995; Yang, 1999) and word sense disambiguation (Schutze, 1998). Theoretical results (Bartell et al., 1995; Ding, 1999; Papadimitriou, Raghavan, Tamaki, & Vempala, 1998; Zha et al., 1998) have also provided some understanding of the effectiveness of LSI.

* Corresponding author.
E-mail addresses: pjrhusbands@lbl.gov (P. Husbands), hdsimon@lbl.gov (H. Simon), chqding@lbl.gov (C. Ding).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.03.006

These LSI studies have, however, mostly used relatively small text collections and simplified document models. In this work we investigate the use of LSI on a larger document collection (TREC). Our initial finding is that on larger text collections, retrieval precision is not enhanced because the LSI mechanism for representing the terms is not sufficient for dealing with the variability in term occurrence. We focus on the norm (the magnitudes) of terms and study the term norm distribution in detail. We propose a term normalization scheme for LSI which improves retrieval precision on the TREC and NPL text collections.

In Section 2 we introduce the concepts of text retrieval and LSI necessary for our work. A short description of our experimental setup is presented in Section 3. Section 4 describes how term occurrence variability affects the SVD and then shows how the decomposition influences retrieval performance. A possible way of improving SVD-based techniques is presented in Section 5 and we conclude in Section 6. A preliminary version of this report appeared in (Husbands, Simon, & Ding, 2001).

2. The vector space model and LSI

In text retrieval (see (Frakes & Baeza-Yates, 1992; Salton & Buckley, 1988; Berry et al., 1995) for treatments of some of the issues), a simple way to represent a collection of documents is with a term-document matrix X with

$$X(i,j) = L(i,j) * G(i)$$

where $L(i,j)$ is a local weighting and $G(i)$ is a global weighting depending on term i. The local weight depends on $\mathrm{tf}(i,j)$, the number of occurrences of term i in document j. In a very simple weighting scheme, one simply uses $X(i,j) = \mathrm{tf}(i,j)$ as the entries of the term-document matrix. However, this scheme is incorrectly dominated by frequent terms.

2.1. Term weighting

Perhaps the most commonly used term weighting scheme is the `tf. idf` weighting scheme. This scheme uses the standard term frequency $\mathrm{tf}(i,j)$, but weighted by the global inverse document frequency (idf). This scheme is specified by

$$L(i,j) = \mathrm{tf}(i,j), \quad G(i) = \mathrm{idf}(i) = \log_2\left(\frac{n}{\mathrm{df}(i)} + 1\right) \tag{1}$$

where n is total number of documents, and $\mathrm{df}(i)$ is the document frequency of term i, the number of documents in which term i occurs. This scheme gives very frequent terms low weight and assigns large weight for infrequent (and hopefully more discriminating) terms.

For comparison purposes we also study the `log. entropy` weighting scheme (Dumais, 1991). In this weighting, the local term weight is the logarithm of the term frequency. The global weighting uses the *entropy* $E(i)$ of term i. This scheme is specified by

$$L(i,j) = \log_2(\mathrm{tf}(i,j) + 1), \quad G(i) = 1 - E(i), \quad E(i) = -\sum_{j=1}^{m} \frac{p_{ij}\log_2(p_{ij})}{\log_2(n)} \tag{2}$$

where $p_{ij} = \frac{\mathrm{tf}(i,j)}{\sum_j \mathrm{tf}(i,j)}$.

Queries (over the same set of terms) are similarly represented. The similarity between document vectors (the columns of term-document matrices) can be found by their inner product. This corresponds to determining the number of term matches (weighted by frequency) in the respective documents. Another commonly used similarity measure is the cosine of the angle between the document vectors. This can be